VERIFICATION OF TRANSLATION

I, Tamotsu Okato, whose address is c/o Hewlett-Packard Japan Ltd. 3-8-13, Takaidohigashi, Suginami-ku, Tokyo, Japan, hereby certify that I am conversant with the Japanese and English languages, and that to the best of my knowledge and belief the accompanying document is a true English translation of Japanese Patent Application No. 2003-105867 (JP 2003-105867).

Signature Tamotsu Okato

Date this 30 day of May , 2008.

[Name of Document] Application for Patent

[Date of submit] March 5th 2003

[Docket Number] 200309904

[Addressee] Patent Office Commissioner

[IPC] G06F 17/21

[Inventor]

[Address or Domicile] 3-29-21 Takaidohigashi, Suginami-ku,
Tokyo
c/o HEWLETT-PACKARD Japan

[Name] Takahiko Kawatani

[Applicant]

[Identification Number] 398038580

[Name] HEWLETT-PACKARD COMPANY

[Agent]

[Identification Number] 100082946

[Patent Attorney]

[Name] Akihiro Onishi

[Official Fees]

[Deposit Account No.] 061492

[Official Fees] 21,000 yen

[List of Attached Documents]

[Name of Document] Specification 1

[Name of Document] Drawings 1

[Name of Document] Abstract 1

[General Power of Attorney Number] 0300115

[Requirement of Proof] Required

[Name of Document] Specification

[Title of the invention] DOCUMENT CLUSTERING METHOD AND APPARATUS

[Claims]

[Claim 1] A clustering method for grouping documents, with respect to a document set including plural documents each having one or plural document segments, on the basis of a relation between the documents, comprising the steps of:

(a) selecting a document having a high commonality to the document set from the document set and making it a cluster;

(b) adding all documents having high similarities to the selected document into the cluster;

(c) adding all documents having high commonalities to the documents existing in the cluster into the cluster repeatedly;

(d) repeating the steps (a), (b) and (c) wherein the step(a) is performed for the document set that is obtained by removing documents having high commonalities to any current clusters from the entire document set and step(b) is performed for the entire document set or for the document set that is obtained by removing documents having high commonalities to any current clusters from the entire document set and step(c) is performed for the entire document set; and

(e) removing, in a case where the cluster overlaps with a redundant cluster, the redundant cluster.

[Claim 2] A method according to claim 1, wherein the step (a) further comprises the steps of:

(a-1) constructing a document segment vector for each of the document segments, in which a value of a component corresponding to a term occurring in the document segment is 1, and another value is 0;

(a-2) obtaining a co-occurrence matrix constructed using the document segment vectors for each of the documents of the document set;

(a-3) obtaining a document frequency matrix from all the documents contained in the document set; and

(a-4) selecting a document having a high commonality to the document set by using the co-occurrence matrix and the document frequency matrix.

[Claim 3] A method according to claim 1, wherein the step (c) further comprises the steps of:

(c-1) obtaining for each term a document frequency ratio from the number of documents containing each term in the input document set and the number of documents containing each term in the cluster;

(c-2) selecting terms or term pairs by using the number of documents containing each term or each term pair in the cluster and information of the document frequency ratio, and obtaining a document commonality by using the selected terms or term pairs; and

(c-3) selecting documents having high commonalities to the selected document on the basis of the document commonality.

[Claim 4] A method according to claim 1, wherein the step (d) further comprises the steps of:

(d-1) obtaining a document commonality between each document and each cluster;

(d-2) obtaining for each cluster the number of documents whose document commonalities to a given cluster is larger than a threshold and whose document commonalities to any other clusters is less than the threshold;

(d-3) deleting, in a case where there are clusters whose number of the documents obtained at (d-2) is less than a threshold, the cluster having the minimum number of the documents obtained at (d-2); and

(d-4) repeating the steps (d-1) to (d-3).

[Claim 5] A clustering method for grouping documents having the same topic with respect to a document set including plural documents each having one or plural document segments, comprising the steps of:

(a) constructing a document segment vector for each of the document segments, in which a value of a component corresponding to a term occurring in the document segment is 1, and another value is 0;

(b) obtaining a co-occurrence matrix constructed from the document segment vector for each of the documents of the document set;

(c) obtaining a document frequency matrix from all documents contained in the document set;

(d) obtaining a common co-occurrence matrix for a set of documents whose document commonalities to any current clusters are less than a threshold;

(e) extracting a document as a seed of a cluster from the set of the documents whose document commonalities to any current clusters are less than a threshold by using the common co-occurrence matrix obtained at the step (d), and constructing an initial cluster byneighbor documents to the seed document of the cluster;

(f) obtaining a common co-occurrence matrix and a document frequency matrix from a document set temporarily belonging to a given cluster;

(g) obtaining information of distinctiveness of each term and each term pair to the given cluster by comparing the document frequency matrixes obtained at the steps (c) and (f);

(h) obtaining a document commonality of each document to the given cluster by using the common co-occurrence matrix obtained at the step (f) and weights of each term and each term pair obtained from the information of the distinctiveness obtained at the step (g) and making a document having a document commonality higher than a threshold belong temporarily to the given cluster;

(i) repeating the steps (f) to (h) until the number of documents temporarily belonging to the given cluster does not increase;

(j) repeating the steps (d) to (i) until the number of documents whose document commonalities to any current clusters are less than a threshold becomes 0, or the number is less than a specific value and is equal tothat of the former repetition;

(k) deciding, on the basis of the document commonality of each document to each cluster, a cluster to which each document belongs; and

(l) checking existence of a redundant cluster, and removing, when the redundant cluster exists, the redundant cluster and again deciding a cluster to which each document belongs.

[Claim 6] A method according to claim 5, further comprising letting M denote the number of sorts of the occurring terms, D_r denote an *rth* document in a document set D consisting of R documents, Y_r denote the number of document segments of the document D_r, and $d_{ry} = (d_{ry1}, \ldots, d_{ryM})^T$ denote a *yth* document segment vector of the document D_r, letter T indicating transposition of a vector, and determining the co-occurrence matrix S^r of the document D_r by:

[Mathematical formula 1]

$$S^r = \sum_{y=1}^{Y_r} d_{ry} d_{ry}^T \quad \ldots (1).$$

[Claim 7] A method according to claim 6, wherein each component of the document frequency matrix of the document set D is the number of documents in which a corresponding component of the co-occurrence matrix of each document in the document set D is not zero.

[Claim 8] A method according to claim 6 or 7, wherein on the basis of a matrix T whose mn component is determined by

[Mathematical formula 2]

$$T_{mn} = \prod_{\substack{r=1 \\ S^r{}_{mn}>0}}^{R} S^r{}_{mn} \quad , \quad \ldots (2)$$

the common co-occurrence matrix of the document set D is given by a matrix T^A whose mn component is determined by

$T^A{}_{mn} = T_{mn}$, $\quad U_{mn} \geq A$,

$T^A{}_{mn} = 0 \quad$ otherwise,

or by a matrix Q^A whose mn component is determined by

$Q^A{}_{mn} = \log(T^A{}_{mn}) \quad T^A{}_{mn} > 1$,

$Q^A{}_{mn} = 0 \quad$ otherwise.

[Claim 9] A method according to claims 6, 7 and 8, letting z_{mm} and z_{mn} be weights for a term m and a term pair m, n, respectively, a document commonality of a document P having the co-occurrence matrix S^P with respect to the document set D is given by

[Mathematical formula 3]

$$com_I(D,P;Q^A) = \frac{\sum_{m=1}^{M} z_{mm} Q^A{}_{mm} S^P{}_{mm}}{\sqrt{\sum_{m=1}^{M} z_{mm} (Q^A{}_{mm})^2} \sqrt{\sum_{m=1}^{M} z_{mm} (S^P{}_{mn})^2}} \quad \ldots (3)$$

or

[Mathematical formula 4]

$$com_q(D,P;Q^A)=\frac{\sum_{m=1}^{M}\sum_{n=1}^{M}z_{mn}Q^A{}_{mn}S^P{}_{mn}}{\sqrt{\sum_{m=1}^{M}\sum_{n=1}^{M}z_{mn}(Q^A{}_{mn})^2}\sqrt{\sum_{m=1}^{M}\sum_{n=1}^{M}z_{mn}(S^P{}_{mn})^2}} \quad \dots (4)$$

or an equation in which the matrix T^A is used instead of the matrix Q^A in equation (3) or equation (4).

[Claim 10] A method according to claims 6, 7, 8 and 9, wherein extraction of the seed document of the cluster and construction of the initial cluster are performed through the steps of:

(a) obtaining a document set commonality of each document in the document set or a quantity of common information by using a common co-occurrence matrix obtained from a set of documents whose document commonalities to any current clustersis less than a threshold;

(b) extracting, as candidates of the seed of a cluster, a specific number of documents whose document set document commonalities or the quantities of the common information obtained at the step (a) is large;

(c) obtaining similarities of the respective candidates of the seed of the cluster to each document in the document set, and obtaining documents whose similarities are larger than a threshold as neighbor documents; and

(d) selecting the candidate whose number of the neighbor documents is the largest among the candidates as the seed of the cluster and making its neighbor documents be the initial cluster.

[Claim 11] A method according to claim 10, wherein the quantity of the common information of the document P having the co-occurrence matrix S^P with respect to the document set D is given by

[Mathematical formula 5]

$$comInfo(D,P)=\sum_{m=1}^{M}\sum_{n=1}^{M}z_{mn}Q^A{}_{mn}S^P{}_{mn} \quad \dots (5)$$

or by an equation in which the matrix T^A is used instead of the matrix Q^A in the numerical equation (5).

[Claim 12] A method according to claims 6 to 11, wherein decision of the information of the distinctiveness of each term and each term pair with respect to the given cluster and the weights is performed through the steps of:

(a) obtaining a ratio of each component of a document frequency matrix obtained from all input documents to a corresponding component of a document frequency matrix obtained from a document set belonging temporarily to the given cluster as a document frequency ratio of each term for each diagonal component and as a document frequency ratio of each term pair for each non-diagonal component;

(b) selecting, in the document set belonging temporarily to the given cluster, a specific number of terms having small document frequency ratios among a specific number of terms having highest document frequencies, and obtaining the average of the document frequency ratios of the selected terms as the average document frequency ratio, alternatively selecting a specific number of terms or term pairs having small document frequency ratios among a specific number of terms or term pairs having highest document frequencies, and obtaining the average of document frequency ratios of the selected terms or term pairs as the average document frequency ratio;

(c) obtaining a value by dividing the average document frequency ratio by the document frequency ratio of each term or each term pair as the information of the distinctiveness of each term or each term pair; and

(d) determining the weight of each term or each term pair by a function having the information of the distinctiveness as a variable.

[Claim 13] A program for an apparatus including a document input part, a document preprocessing part, a document information processing part, and an output part and for grouping documents, with respect to an input document set including plural documents each having one or plural document segments, on the basis of a relation between the documents, the program causing operations of:

(a) a cluster selection part for selecting a document having the highest commonality to the document set from the document set and making it be a cluster;

(b) an merging part for adding all documents having high similarities to the selected document into the cluster;

(c) a cluster growing part for adding all documents having high commonalities to the documents existing in the cluster into the cluster repeatedly;

(d) a control part for repeating the steps (a), (b) and (c) for the document set that is obtained by removing documents having high commonalities to any current clusters from the entire document set;; and (e) a removing part for removing, in a case where the cluster overlaps with a redundant cluster, the redundant cluster.

[Claim 14] A clustering apparatus comprising a document input part, a document preprocessing part, a document information processing part, and an output part and for grouping documents, with respect to a document set including plural documents each having one or plural document segments, on the basis of a relation between the documents, the apparatus comprising:

(a) means for selecting a document having a high commonality to the document set from the document set and making it a cluster;

(b) means for adding all documents having high similarities to the selected document into the cluster;

(c) means for adding all documents having high commonalities to the document existing in the cluster into the cluster repeatedly;

(d) means for repeating steps of the means (a), (b) and (c) for the document set that is obtained by removing documents having high commonalities to any current clusters from the entire document set;; and

(e) means for removing, in a case where the cluster overlaps with a redundant cluster, the redundant cluster.

[Detailed explanation of the invention]

[0001]

[Field of the invention]

The present invention relates to a natural language processing including document clustering, and facilitates information extraction from documents by improving the performance of the processing.

[0002]

[Prior art]

The document clustering is a technique for dividing an inputted document set into some groups according to the contents of documents. The clustering technique has been studied for a long time, and methods hitherto devised are systematically introduced in "Foundations of Statistical Natural Language Processing" (The MIT Press, 1999) written by C.D. Manning and H. Schutze. Roughly speaking, there are two kinds of approaches in clustering. One is soft clustering for obtaining a probability that each document belongs to each cluster, and the other hard clustering for determining whether or not each document belongs to each cluster. The latter is further divided into a hierarchical approach and a non-hierarchical one. The hierarchical method is further divided into a bottom-up approach and a top-down one. In the former, as an initial state, each document becomes a seed of a cluster, and a processing of merging closest clusters is repeated. By this operation, a document set is expressed in a tree structure. As the methods of measuring the degree of closeness between clusters, that is, similarity, a single link method, a complete link method, or a group average method are well known. In any of these, calculation is performed on the basis of the similarity between two documents. In the top down approach, from an initial state where all documents belong to one cluster, a processing is repeated in which for example, in the case where the lowest similarity in all document pairs in one cluster is less than a threshold value, the cluster is divided.

[0003]

In the non-hierarchical method, a previously determined number of clusters are constructed so as to satisfy some standard. Typical processing of this type is as follows.

step 1: a step of randomly selecting a specified cluster number of documents and making them centers of the respective clusters,

step 2: a step of obtaining the closeness between each document and the center of each cluster and making each document belong to the closest cluster,

step 3: a step of obtaining the center of each cluster by averaging document vectors belonging to each cluster, and

step 4: a step of carrying out the processing of the step 2 and if the cluster to which each document belongs is not changed, the procedure is ended, and if not, it returns to the step 3.

[0004]

[Problems solved by the invention]

The conventional document clustering technique has two serious problems. One of them is a problem about the number of clusters to be obtained. In the document clustering, the number of the clusters to be obtained must be the same as the number of topics stated in documents of an inputted document set. As described above, in the bottom-up hierarchical clustering processing, each cluster starts from the state including one document, the processing of merging closest clusters is repeated, and all documents finally belong to one cluster. Accordingly, in order to obtain clusters whose number is same as the number of topics, it becomes necessary to stop merging of clusters. This can be realized by not performing merging of a cluster pair having similarity lower than a threshold value in the merging processing of clusters. However, it is actually a difficult how to determine the threshold value. If the threshold value is inadequate, a correct number of clusters can not be obtained. Similarly, in the top-down clustering processing, if a cluster is not divided in the case where the lowest similarity in all document pairs in one cluster is higher than a threshold value, clusters whose number is same as the number of topics ought to be obtained in principle.

[0005]

However, also in this case, it is a difficult problem how to determine the threshold value. Besides, in the non-hierarchical clustering, the user is required to input information in advance as to the number of clusters into which a given document set is divided. However, it is impossible to accurately give the information of the number of clusters without previous knowledge of the input document set. As stated above, it is a difficult problem to obtain a correct number of clusters from the input document set. Although the performance has been improved by Liu et al's attempt to correctly infer the number of clusters in non-hierarchical clustering, it is not perfect (X. Liu, Y.

Gong, W. Xu and S. Zhu, Document Clustering with Cluster Refinement and Model Selection Capabilities. In Proceedings of the 25th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp. 191 - 198. Tampere, Finland, August, 2002).

[0006]

The second problem is a problem of accuracy in clustering. This is a problem as to whether documents belonging to the same cluster describe the same topic. In clustering processing, in general, a document is expressed by a vector in which each component is according to existence of each term in the document or an occurrence frequency. So, the similarity between two clusters is obtained on the basis of cosine similarity between two vectors of documents belonging to different clusters, and the distance between a certain document and a cluster is obtained on the basis of a distance (for example, Euclidean distance) between the vector of the document and the average vector of documents belonging to the cluster. In the conventional clustering processing, when the cosine similarity or the Euclidean distance is obtained, a vector obtained in each document is usually used without verifying what term is important for the cluster. Thus, the existence of a term or a term pair which is not essential to each cluster can have an influence on the accuracy of the clustering.

[0007]

[Means to solve the problems]

Since document clustering groups documents according to a topic described in each document, documents (called cluster document set) belonging to one cluster ought to describe the same topic. Accordingly, the cluster document set ought to have some commonality. Besides, each topic ought to have terms or term pairs distinctive to the topic, which frequently occurs in the topic and seldom occurs in other topics. Accordingly, there ought to be differences in occurrence tendencies of terms or term pairs among clusters. In view of the above, in this invention, in order to increase the accuracy of the clustering, the following means are introduced in a process of clustering.

(A) Common information of a given cluster document set is extracted, and closeness (document commonality) of each document to the given cluster is obtained using common information.

(B) Terms and term pairs not distinctive to the given cluster are detected, and the influence of those is removed in the calculation of the document commonality.

[0008]

In the conventional hierarchical processing, merging or dividing of clusters are repeated many times. In the conventional non-hierarchical processing, members of clusters are interchanged many times. In such a situation, it is difficult to detect common information of the respective clusters, or terms and term pairs not distinctive to the clusters. Then, in this invention, the following is adopted as the whole procedure of clustering.

[0009]

Step 1: Candidates of a cluster seed are detected in the first iteration from all documents, and in the second or subsequent iteration from documents in which document commonalities to any current clusters are less than a threshold.

Step 2: First, with respect to each candidate, similarities to all documents are obtained, and documents having similarities higher than a threshold are extracted as neighbor documents. The candidate document which has the greatest number of neighbor documents is selected as the seed of the cluster, and the cluster is constructed by the set of its neighbor documents.

Step 3: A document commonality of each document to current clusters are obtained, and documents having document commonalities higher than a threshold are made to temporarily belong to the cluster, so that the cluster grows up. If the number of documents temporarily belonging to the cluster becomes constant, the procedure proceeds to step 4. If not, this step is repeated.

Step 4: If a termination condition is satisfied, the procedure proceeds to step 5. If not, it returns to the step 1 and continues.

Step 5: With respect to each document, a document commonality to each cluster is obtained, and each document is judged to belong to a cluster to which the document commonality is the highest.

Step 6: It is detected whether two or more clusters overlap and correspond to one topic. If such a cluster exists, it is deleted as a redundant cluster, and a cluster to which each document belongs is again obtained.

[0010]

In the above (A) clustering procedure, the calculation of the document commonality using the common information, and the detection of the term or the term pair not distinctive to the given cluster are carried out at steps 3 and 5. With respect to the former (A), the common information is extracted from the document temporarily belonging to the given. With respect to the extraction and use of the common information, a method disclosed in Japanese Patent Application No. 2002-326157 can be adopted. The basis idea is as follows. Now, it is assumed that a given cluster is composed of R documents, and a sentence group composed of R sentences is constructed by extracting one sentence from the respective documents. The sentence groups are constructed for all possible combinations of sentences. The total number of such sentence groups becomes equal to the number of the product of the numbers of the sentences of the respective documents. Here, in a given sentence group, a term occurring in A sentences among the R sentences is defined as a common term, and a sentence constructed by the common terms is called a common sentence. Here, it is assumed that common sentences are constructed for all the sentence groups, and that a set of the common sentences is constructed. It is conceivable that the set of the common sentences as stated above represents the content of the common topic of the given cluster. Accordingly, if a similarity between each document and the common sentence set can be obtained, it will represent the closeness of each document to the common topic of the given cluster.

[0011]

While, detection of terms and term pairs which are not distinctive to the given cluster (B) is performed based on the following ideas. Consideration will be given to a process of growth of a given cluster whose seed document has

topic i. It is assumed that the number of documents describing topic i is c_0 in the whole document set, and c in a document set of the given cluster. Besides, it is assumed that the number of documents containing a term m is U^0_{mm} in the whole document set, and U_{mm} in the document set of the given cluster. If term m is distinctive for topic i, since most documents that term m occurs has topic i, the following relationship ought to be satisfied:

[0012]

[Mathematical formula 6]

$$U^0_{mm}/U_{mm} \approx c_0/c$$

[0013]

and when it is not distinctive, since the term m occurs frequently in documents of topics other than topic i, the following relationship ought to be satisfied:

[0014]

[Mathematical formula 7]

$$U^0_{mm}/U_{mm} > c_0/c$$

[0015]

Accordingly, if c_0/c can be obtained by a proper method, it becomes able to judge whether or not term m is distinctive to topic i. U^0_{mm}/U_{mm} is called a document frequency ratio of term m. Among a specific number of terms having highest frequencies in the document set of the given cluster, in this invention, a specific number of terms having small document frequency ratios are assumed to be distinctive to topic i, and the average c' of the document frequency ratios of these terms is regarded as the predicted value of c_0/c. Eventually, when α is a parameter, it can be judged that term m satisfying the following equation is not distinctive to topic i.

[0016]

[Mathematical formula 8]

$$U^0_{mm}/U_{mm} > \alpha c'$$

[0017]

Similarly, it is assumed that the number of documents containing terms m, n is U^0_{mn} in the whole document set, and U_{mn} in the document set of the given cluster, it is able to judge that term pair m, n satisfying the following equation is not distinctive to topic i.

[0018]

[Mathematical formula 9]

$U^0_{mn}/U_{mn} > \alpha c'$

[0019]

With respect to the document commonality, in order to reduce the influence of the terms and the term pairs not essential to the given cluster, it is appropriate that the term and the term pair judged not to be distinctive to topic i are not used for the calculation of the document commonality between each document and the document set of the given cluster. Alternatively,

[0020]

[Mathematical formula 10]

$c'/(U^0_{mm}/U_{mm})$

[0021]

[Mathematical formula 11]

$c'/(U^0_{mn}/U_{mn})$

[0022]

can be used as weights of term m and term pair m, n respectively in calculation of the document commonality. By this, the document commonality comes to have a large value for the document describing topic i. As the result, the improvement of accuracy of the clustering can be expected.

[0023]

In the whole procedure of clustering, a processing is repeatedly carried out in which first, one document is extracted as a seed of a cluster, and then the seed grows up by detecting and merging the documents describing the same topic as the seed. Accordingly, if the number of seed documents is just coincident with the number of topics in the input document, a correct number of clusters can be obtained. Even if two seed documents are detected for the same topic at the step 1, since the redundant cluster is detected and removed at step 6, the correct number of clusters can be obtained. If a seed document is not detected for some topic at step 1, the number of clusters becomes short. Such a situation will occur when documents having the topic to be detected have high document similarities to an existing cluster of other topic and are merged to the cluster. However, in this invention, since the accuracy

of the clustering increased by adopting means A) and B), a possibility that documents having a different topic are mixed is low, and there hardly occurs a situation in which the number of obtained clusters becomes short.

[0024]

[Embodiments]

Fig. 1 is a block diagram showing the outline of the present invention. Numeral 110 denotes a document input block; 120, a document preprocessing block; 130, a document information processing block; and 140, an output block. A document set to be processed is inputted to the document input block 110. In the document preprocessing block 120, term detection, morphological analysis, document segment division of an inputted document are performed. A document segment will be described. The document segment is an element composing a document, and its basic unit is a sentence. In the case of an English sentence, since the sentence is ended with a period, and a space follows after that, so that cutout of the sentence can be easily performed. As another document segmentation method, there is a method in which in a case where a sentence is complexed , it is divided into a principle clause and a subordinate clause, or a method in which plural sentences are collected into a document segment so that the number of terms almost becomes the same, or a method in which a document is divided into segments having the same number of terms, from the head thereof and irrespective of sentences. Block 130 performs information processing and processing directly relating to the clustering, such as detection of a seed document, calculation of document set commonalities between all documents and a given cluster, and detection of terms and term pairs not distinctive to each cluster. Block 130 will be described in detail later. The output block 140 outputs the result obtained in the document information processing block 130 to an output device such as a display.

[0025]

Fig. 2 shows an embodiment of this invention in which clustering is performed to a given document set. The method of this invention can be carried out by running a program incorporating this invention on a general-purpose computer. Fig. 2 is a flowchart of a computer in a state

where such a program runs. A block 21 indicates document set input, a block 22 indicates document preprocessing for all documents, a block 23 indicates document set information extraction processing for all documents, a block 24 indicates extraction of a seed document of a cluster and construction of an initial cluster, a block 25 indicates growing processing of a cluster, a block 26 indicates extraction of remaining documents, a block 27 indicates termination condition checking, a block 28 indicates document set information extraction processing for remaining documents, a block 29 indicates decision of cluster member, and a block 30 indicates extraction and removal of redundant clusters. Hereinafter, an embodiment will be described while an English document is used as an example.

[0026]

First, a document set as an object is inputted at the document set input 21. In the document preprocessing 22, a preprocessing such as term detection, morphological analysis, document segment dividing, and document segment vector construction are performed for each input document. As term detection, words, numerical expressions, symbol series, and the like are detected from each input document. Here, a word, a symbol series and the like are generically called a term. In the case of the English writing, since the notation method in which the terms are spaced apart is established, the detection of the terms is easy. Next, in the morphological analysis, morphological analysis such as part of speech tagging to terms is performed for each input document. In the document segment dividing, document segmentation is performed for each input document. In the document segment vector construction, first, the dimensions of a vector to be constructed and the correspondence between each component and each term are determined from terms occurring in the whole document. It is not necessary to make components of the vector correspond to all terms occurring in the document, and by using the result of the processing of part of speech tagging, the vector may be constructed by using, for example, only terms judged to be nouns and verbs. Next, the document segment vector is constructed in which only components corresponding to terms occurring in each document segment are 1, and the others are 0.

[0027]

In the document set information extraction processing 23 for all documents, data used in the clustering processing stage are obtained from each document and the whole input document set. The data to be obtained are a co-occurrence matrix of each document, a co-occurrence matrix (common co-occurrence matrix) of the document set, and a document frequency matrix of the whole input document set. The co-occurrence matrix of each document is a matrix reflecting occurrence frequencies of terms, and co-occurrence frequencies of term pairs. The description will be continued on a case where a sentence is a document segment. Letting M denote the number of kinds of the occurring terms, D_r denote an r-th document in a document set D consisting of R documents, Y_r denote the number of sentences of the document D_r, and $d_{ry} = (d_{ry1}, ..., d_{ryM})^T$ denote a y-th sentence vector. Since the sentence vector d_{ry} is a binary vector, d_{rym} denotes the existence or absence of the m-th term. Let S^r be the co-occurrence matrix of the document D_r. S^r is given by

[0028]

[Mathematical formula 12]

$$S^r = \sum_{y=1}^{Y_r} d_{ry} d_{ry}^T, \quad \ldots (1)$$

[0029]

where T denotes vector transpose.

As is apparent from equation (1), the mn components of S^r is given by

$$S^r{}_{mn} = \sum_{y=1}^{Y_r} d_{rym} d_{ryn}.$$

Therefore, $S^r{}_{mm}$ represent the occurrence counts of sentences in which term m occur and $S^r{}_{mn}$ represent the co-occurrence counts of sentences in which term and n co-occur. If each term dooes not occur twice r more in each sentence, $S^r{}_{mm}$ represent the occurrence frequency of term m in document D_r. Next, matrix T whose mn component is defined as follows is obtained.

[0030]

[Mathematical formula 13]

$$T_{mn} = \prod_{r=1, S^r{}_{mn}>0}^{R} S^r{}_{mn}$$

[0031]

Further, a matrix U^0 storing document frequencies of each term and term is obtained. U^0_{mm} and U^0_{mn} denote the number of documents in which the term m occurs, and the number of documents in which the terms m and n co-occur, respectively. By using the matrix T and U^0 as stated above, the common co-occurrence matrix T^A is obtained. A mn component of the common co-occurrence matrix T^A is determined as follows.

$T^A_{mn} = T_{mn}$ $U^0_{mn} \geq A$,

$T^A_{mn} = 0$ otherwise.

"A" denotes a threshold that is experimentally determined.

[0032]

Besides, a matrix Q^A in which a mn component is given below is defined, and may be used as the common co-occurrence matrix.

$Q^A_{mn} = \log(T^A_{mn})$ $T^A_{mn} > 1$,

$Q^A_{mn} = 0$ otherwise.

[0033]

In the block 24 of the extraction of a seed document of a cluster and the construction of the initial cluster, processings corresponding to the steps 1 and 2 are performed. Here, Let D' be a set of documents whose document commonalities to any current clusters are less than a threshold. The document set D' is a set of documents having a high possibility that they do not belong to any current clusters. The common co-occurrence matrix T^A, Q^A, and the document frequency matrix U are calculated on the basis of all documents D at the first iteration, and are calculated on the basis of the document set D' at the second and subsequent iteration. It is desirable that the seed document of a cluster is the dominant document in the topic that the document describes. In this invention, on the assumption that the dominant document in a document group of the most dominant topic in D' has a high document commonality to D', the document commonality between each document in the document set D' and the document set D' is obtained, and the documents having high document commonalities are selected as candidates of the seed of the cluster. Let S^P be a common co-occurrence matrix for an arbitrary document P. The document commonality between the

document P and the document set D', for example, the following can be obtained as follows.

[0034]

[Mathematical formula 14]

$$com_q(D', P; Q^A) = \frac{\sum_{m=1}^{M}\sum_{n=1}^{M} Q^A{}_{mn} S^P{}_{mn}}{\sqrt{\sum_{m=1}^{M}\sum_{n=1}^{M} (Q^A{}_{mn})^2}\sqrt{\sum_{m=1}^{M}\sum_{n=1}^{M} (S^P{}_{mn})^2}} \quad \cdots (2)$$

[0035]

In the equation (2), the matrix T^A can also be used instead of the matrix Q^A. Besides, in the equation (2), in order to reduce the influence of terms common to plural topics, diagonal components of the co-occurrence matrix and the common co-occurrence matrix may not be used because individual terms tend to be shared in plural topics more easily than term pairs.

[0036]

The candidates of the seed document of the cluster are obtained by calculating the document commonalities to all documents in D' using the equation (2), and by selecting a specific number of documents having high document commonalities. Next, the cluster seed document extraction will be described. First, for each candidate document, the similarities to all documents in D' are obtained. Next, for each candidate document, documents having greater similarities than a preset threshold are obtained as neighbor documents of each candidate document. The document in which the number of neighbor documents is the largest is selected from the candidate documentsas the cluster seed document. The initial cluster is given by the neighbor documents of the seed document.

[0037]

In the growing processing 25 of the cluster, a cluster grows up by merging documents having high commonalities to the cluster. Fig. 3 is a block diagram of such a. Reference numeral 31 denotes construction of document frequency matrix; 32 construction of common co-occurrence matrix; 33 distinctiveness calculation for each term and term pair; 34 document commonality calculation; 35 decision of cluster member; and 36 termination condition checking.

[0038]

In the block 31 of the construction of the document frequency matrix and in the block 32 of the construction of the common co-occurrence matrix, processings equivalent to the document frequency matrix construction processing and the common co-occurrence matrix construction processing in the block 23 of Fig. 2 are performed to the set of documents which are currentmembers of the given. Let U be the document frequency matrix obtained at 31. Let T^A and Q^A be the common co-occurrence matrix obtained at 32 and its modified one, respectively. In the block 33, the distinctiveness and the weight are determined for each term and term pair. First, as described before, U^0_{mm}/U_{mm} is obtained as the document frequency ratio of the term m, and among a specific number of terms having high document frequencies a specific number of terms having small document frequency ratios are selected. They are assumed to be distinctive terms of the given cluster. Next, the document frequency ratios of these terms are averaged. Let c' be the average. Distinctiveness v_{mm} of the term m, and distinctiveness v_{mn} of the term pair m, n are determined by the followings.

[0039]

[Mathematical formula 15]

$$v_{mm} = c' / (U^0_{mm} / U_{mm})$$

[0040]

[Mathematical formula 16]

$$v_{mn} = c' / (U^0_{mn} / U_{mn})$$

[0041]

Alternatively, the average document frequency ratio may be obtained by using both the distinctive term pair and the distinctive term. In this case, U^0_{mn}/U_{mn} is obtained as the document frequency ratio of the term pair m, n when m does not equal n, and is obtained as the document frequency ratio of the term m when m equals n. Among a specific number of terms and term pairs having the highest document frequencies, a specific number of terms or term pairs having small document frequency ratios are selected. They are regarded as the distinctive terms or term pairs of the given cluster. Next, the document frequency ratios of these terms and term pairs are averaged. Let c' be the average.

[0042]

Let z_{mm} and z_{mn} be the weights of the term m and the term pair m, n, respectively. These are determined by using a weight deciding function f(x) as follows.

[0043]

[Mathematical formula 17]

$$z_{mm} = f(v_{mm})$$

[0044]

[Mathematical formula 18]

$$z_{mn} = f(v_{mn})$$

[0045]

There can be considered many functions for f(x). For example, the following can be used.

f(x) = x,

or

$f(x) = x^2$,

or

f(x) = 1 if x>threshold,

f(x) = 0 otherwise.

[0046]

In the document commonality calculation 34, the document commonalities to the given cluster are calculated for all input documents. Let S^P be the co-occurrence matrix of document P. The document commonality of the document P to document set D can be obtained by

[0047]

[Mathematical formula 19]

$$com_i(D,P;Q^A) = \frac{\sum_{m=1}^{M} z_{mm} Q^A{}_{mm} S^P{}_{mm}}{\sqrt{\sum_{m=1}^{M} z_{mm} (Q^A{}_{mm})^2} \sqrt{\sum_{m=1}^{M} z_{mm} (S^P{}_{mn})^2}},$$

[0048]

or

[0049]

[Mathematical formula 20]

$$com_q(D,P;Q^A) = \frac{\sum_{m=1}^{M}\sum_{n=1}^{M} z_{mn} Q^A{}_{mn} S^P{}_{mn}}{\sqrt{\sum_{m=1}^{M}\sum_{n=1}^{M} z_{mn} (Q^A{}_{mn})^2}\sqrt{\sum_{m=1}^{M}\sum_{n=1}^{M} z_{mn} (S^P{}_{mn})^2}},$$

[0050]

where D represent a document set of the given cluster, in the above equation, the matrix T^A can also be used instead of the matrix Q^A.

[0051]

In the block 35 of the determination of the cluster member, documents having the higher document commonalities than a specific value are selected as the temporal members of the given cluster.

[0052]

In the termination condition checking 36, it is checked whether or not the growing processing of the given cluster is terminated. First, at the first iteration, that is, when the procedure first reaches the block 36, it returns to the block 31 unconditionally and the processing is repeated. At the second or subsequent iteration, the number of documents in the given cluster obtained at the block 35 is counted, and in the case where it is not equal to that in the former iteration, the procedure returns to 31 and the processing is repeated. If equal, the document commonalities to the given cluster are kept for all input documents, and the growing processing of the given cluster is terminated.

[0053]

Returning to Fig. 2, the description will be continued. In the block 26 of extraction of remaining documents, on the basis of the document commonality of each document to all current clusters, documents whose document commonalities to any clusters are less than a threshold are extracted as remaining documents.

[0054]

In the termination condition checking 27, on the basis of the number of remaining documents, it is checked whether or not a series of processings from the seed extraction to the growing is terminated. For example, in the case where the number of remaining documents is less than a threshold and is equal to the number of remaining documents in the former iteration, the procedure proceeds to the block 29. If such a condition is not satisfied, the

procedure proceeds to the block 28 and the processing equivalent to the block 23 is performed to the remaining document set.

[0055]

In the block 29 of the decision of the cluster member, the cluster that each document belongs to is determined. This can be carried out by using the information of the document commonality to each cluster obtained for each document in Fig. 3 and by making each document belong to the cluster to which the document commonality is the highest.

[0056]

In the block 30 of the detection and removal of redundant clusters, it is checked whether or not a redundant cluster exists, and in the case where it exists, it is removed. The redundant cluster occurs when two or more clusters are obtained for one topic. In such a case, the document describing the topic will have large document commonalities to two or more clusters, and the two or more clusters will overlap with each other. For the detection of the redundant clusters, first, the document commonalities to all obtained clusters are obtained for all documents, and next, the number of documents whose document commonalities to a given cluster are larger than a threshold and whose document commonalities to any other cluster are less than the threshold is obtained. In the case where the given cluster does not overlap with any other clusters, the number of such documents becomes equal to the number of documents having the higher document commonalities than the threshld to the given cluster. On the other hand, in the case where the given cluster overlaps with another cluster, it becomes the number of documents not overlapping with the cluster, that is, the number of documents belonging to only the given cluster. The number of documents as stated above can defined as importance of each cluster. In the case of Fig. 4A, for example, the importance of cluster 1 is the number of documents belonging to cluster 1 This is the case for cluster 2. In the case where the given cluster partially overlaps with the other cluster, the importance becomes the number of documents not overlapping with the other cluster. That is, with respect to the cluster 1, the importance is represented by the number of documents contained in a portion indicated by "c" of Fig. 4B. With respect to the cluster

2, the importance is represented by the number of documents contained in a portion shown by "d" of Fig. 4B. In the case where the importance of one cluster is smaller than a specific value, even if the number of documents belonging to the cluster is large, it is regarded as the redundant cluster and is removed. If plural such clusters exist, the cluster having the lowest cluster importance is first removed. With respect to the remaining clusters, calculation of cluster importance is performed again, and the cluster having the lowest cluster importance is removed. The processing like this is repeated until no redundant clusters exists. In the case where the redundant cluster removal is performed, the determination of each cluster member is performed again.

[0057]

[Effects of the invention]

Here, in order to explain the effect of this invention, experimental results along the embodiment of Figs. 2 and 3 will be shown. TDT2 is used as the corpus. The corpus TDT2 is a set of news stories relating to 100 events between January of 1998 to June thereof, and is gathered from six news sources. For comparison with the result of non-hierarchical clustering performed using TDT2 by Liu et al. (X. Liu, Y. Gong, W. Xu and S. Zhu, Document Clustering with Cluster Refinement and Model Selection Capabilities. In Proceedings of the 25th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp. 191 - 198. Tampere, Finland, August, 2002), experimental results using the same data in experiments in Liu et al. will be shown. The data are a set of news stories relating to 15 events gathered from ABC, CNN and VOA is made an experimental object. Table 1 shows the details of those.

[0058]

[Table 1]

Event ID	Content of Each Event	Number of Documents			
		ABC	CNN	VOA	Total
01	Asian Economic Crisis	27	90	289	406
02	Monica Lewinsky Case	102	497	96	695
13	1998 Winter Olympic	21	81	108	210
15	Current Conflict with Iraq	77	438	345	860
18	Bombing AL Clinic	9	73	5	87
23	Violence in Algeria	1	1	60	62
32	Sgt. Gene McKinney	6	91	3	100
39	India Parliamentary Election	1	1	29	31
44	National Tobacco Settlement	26	163	17	206
48	Jonesboro Shooting	13	73	15	101
70	India, A Nuclear Power?	24	98	129	251
71	Israeli-Palestinian Talks	5	62	48	115
76	Anti-Suharto Violence	13	55	114	182
77	Unabomer	9	66	6	81
86	GM Strike	14	83	24	121

[0059]

Table 2 shows 14 data sets used in the experiment, and the accuracy of clustering of the proposed method and the method of Liu et al to those. The results of the method of Liu et al. are referred from the paper of Liu et al. In this invention, when the event to which a certain document belongs coincides with the event of the seed document of the cluster, the result of the clustering is regarded as correct. Besides, a document whose document commonality to all clusters is 0 is regarded as erroneous. The accuracy is obtained by a ratio of the number of correctly clustered documents to the number of all documents. In the method of Liu et al., after non-hierarchical clustering is performed on the basis of a Gaussian mixture model, distinctive terms of each cluster are obtained, and the result is corrected by voting of the distinctive terms. In Table 2, ABC-01-02-15 of the test data means documents gathered from ABC and having event IDs belonging to 01, 02 and 15. From Table 2, the number of data sets having high accuracy in this invention is

larger than that in the method of Liu et al., and it is understood that this invention outperforms.

[0060]

[Table 2]

Number	Data Set	Method of Liu et al.	This Invention
1	ABC-01-02-15	1.0000	0.9806
2	ABC-02-15-44	0.9902	0.9805
3	ABC-01-13-44-70	1.0000	1.0000
4	ABC-01-44-48-70	1.0000	1.0000
5	CNN-01-02-15	0.9756	0.9932
6	CNN-02-15-44	0.9964	0.9964
7	VOA-01-02-15	0.9896	0.9986
8	VOA-01-13-76	0.9583	0.8943
9	VOA-01-23-70-76	0.9453	0.9206
10	VOA-12-39-48-71	0.9898	1.0000
11	VOA-44-48-70-71-76-77-86	0.8527	1.0000
12	ABC+CNN-01-13-18-32-48-70-71-77-86	0.9704	0.9917
13	CNN+VOA-01-13-48-70-71-76-77-86	0.9262	0.9500
14	ABC+CNN+VOA-44-48-70-71-76-77-86	0.9938	1.0000

[0061]

The number of extracted clusters was correct for all the data of Table 2 in this invention.

Besides, also with respect to 12 data sets listed in the paper of Liu et al., the number of extracted clusters was correct in this invention. On the other hand, in the method of Liu et al., the number of extracted clusters was incorrect for three data sets among 12 data sets. Table 3 shows the results of the method of Liu et al. and this invention.

[Table 3]

Test Data	Number of Clusters To Be Obtained	Testing Results by Liu et al.	Testing Results by This Invention
ABC-01-03	2	2	2
ABC-01-02-15	3	3	3
ABC-02-48-70	3	2	3
ABC-44-70-01-13	4	4	4
ABC-44-48-70-76	4	4	4
CNN-01-02-15	3	4	3
CNN-01-02-13-15-18	5	5	5
CNN-44-48-70-71-76-77	6	5	6
VOA-01-02-15	3	3	3
VOA-01-13-76	3	3	3
VOA-01-23-70-76	4	4	4
VOA-12-39-48-71	4	4	4

[0062]

As described above, according to this invention, a correct number of clusters can be extracted from an input document set, and each document can be assigned to a cluster with high accuracy. Therefore, the efficiency of information acquisition by the user can be significantly improved.

[Brief Description of the Drawings]

[Fig.1] Fig. 1 is a block diagram showing the outline of this invention.

[Fig.2] Fig. 2 shows a procedure from the stage where a document set is inputted to the stage of determination of clusters to which each document belongs.

[Fig.3] Fig. 3 shows, with respect to a cluster, a procedure of growth from the initial cluster.

[Fig.4] Figs. 4A and 4B explains the importance of clusters for deleting a redundant cluster.

[Brief Explanation of the sign]

110: document input block

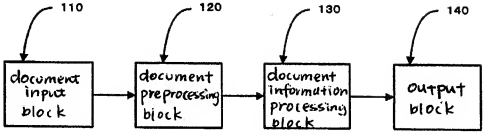
120: document preprocessing block

130: document information processing block

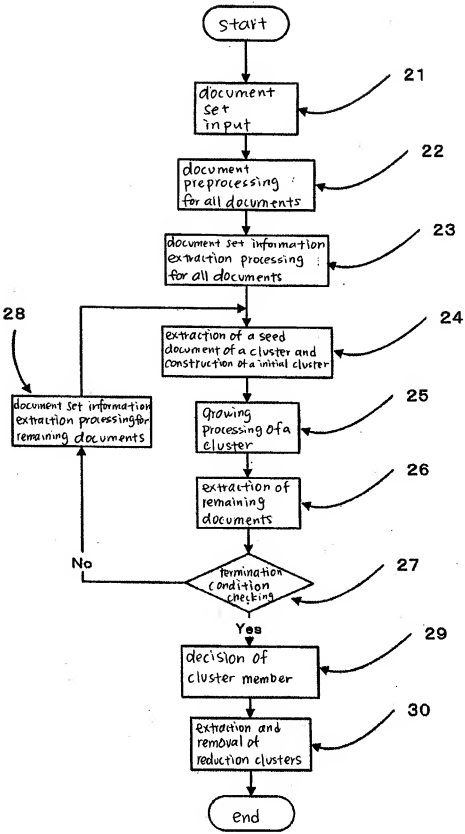
140: output block

[Name of Document] Drawings

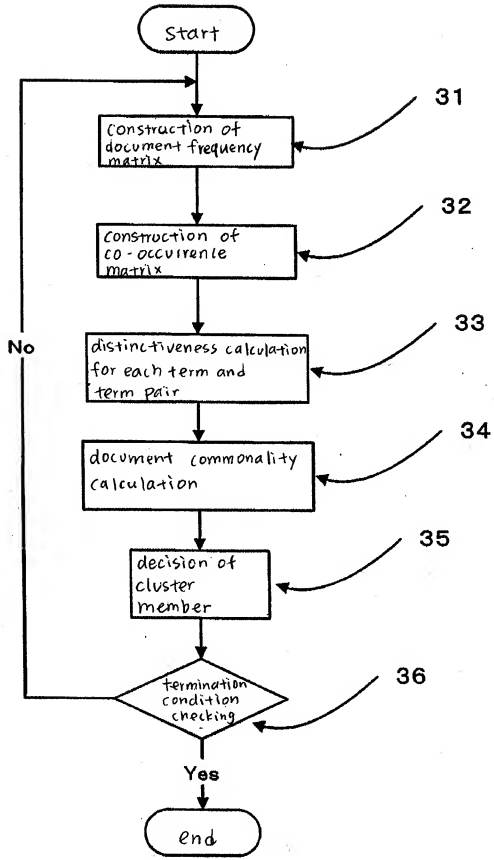
[Fig.1]



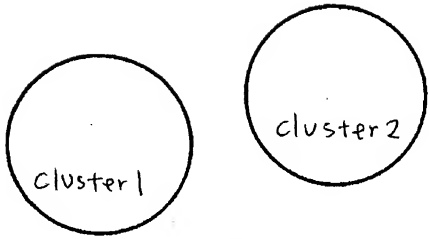
[Fig.2]


start
document set input
21
document preprocessing for all documents
22
document set information extraction processing for all documents
23
28
extraction of a seed document of a cluster and construction of a initial cluster
24
document set information extraction processing for remaining documents
growing processing of a cluster
25
extraction of remaining documents
26
No
termination condition checking
27
Yes
decision of cluster member
29
extraction and removal of reduction clusters
30
end

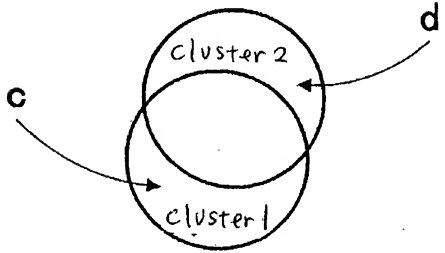
[Fig.3]


Start
31
construction of document frequency matrix
32
construction of co-occurrence matrix
33
distinctiveness calculation for each term and term pair
No
34
document commonality calculation
35
decision of cluster member
36
termination condition checking
Yes
end

[Fig.4]



(a)



(b)

[Name of Document] Abstract of the Disclosure

[Problems]

In document clustering, obtaining a correct number of clusters and accurate assignment of each document to a correct cluster have been not completely solved problems.

[Mean to solve the problems]

In the document clustering, since documents describing the same topic are grouped, a document group belonging to the same cluster ought to have some commonality. Besides, each topic has distinctive terms or term pairs. In this invention, attention is paid to these points, and when the closeness of each document to a given cluster is obtained, common information of the given cluster is extracted and used while the influence of terms or term pairs not distinctive to the given cluster is excluded,.

[Elected drawings] Figure 1